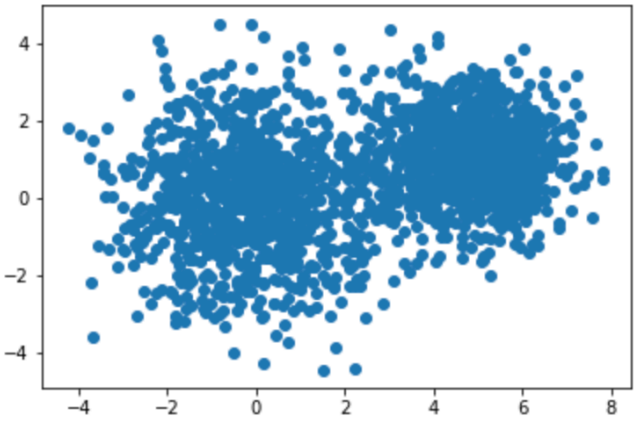
IE 529

Stats of Big Data & Clustering

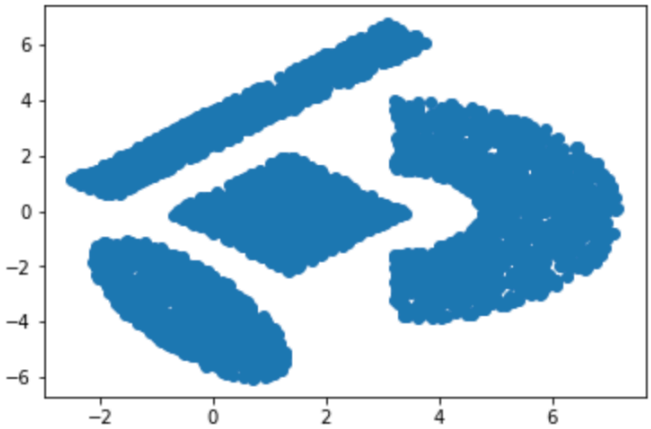
Computations: 2

1. Include a scatter plot for each set of data, as is, i.e., with no clusters identified yet.

Dataset 1



Dataset 2



2. Submit your pseudo-codes for the Lloyds/K-means algorithm, the GreedyKcenters and Single Swap algorithms, and the Spectral Clustering algorithm, each on a separate sheet of paper. State clearly whether you are considering normalized or un-normalized Spectral Clustering.

(1) Pseudo-code for Lloyds algorithm

- Initialize K centroids for the dataset given user-based K.
- Compute the distance (L2-norm) between sample and centroids, assign the centroids index to sample with the least distance between them.
- While converge condition:
  - Re-compute the centroids by compute the centroids of grouping samples.
  - Re-assignment the index of centroids of samples
- End while

(2) Pseudo-code for GreedyKCenters

- Initialize first centroids for the sample data, randomly
- While # centroids < K:
  - Let X_l belongs to the subset of sample data removing the set of centroids which maximize the distance of (X_l, C)
  - Set the new set of centroids be the set union with X_l
- End while

(3) Pseudo-code for Single Swap

- Initialize set C for K centroids
- While exist C not empty and X_i belongs to the subset of X removing C:
  - Make dist(X \ C) < dist(C)
  - Set $C = C \setminus m_j \cup X_i$
- End while

(4) Pseudo-code for Spectral Clustering

Un-normalized

- Initialize the similarity/weighted matrix W
- Initialize degree matrix D
- Initialize Laplacian matrix L = D – W
- Compute eigen/ spectral decomposition of L: U is the matrix eigenvectors, Λ is the diagonal matrix of eigenvalues
- Consider the first K eigenvectors of U (Y) -> associated with K smallest eigenvalues
- Perform K-Means on Y

3. For your K-means algorithm by itself, and for your Spectral Clustering algorithm, present the output (clustering) results for BOTH sets of test data given to you (see below).

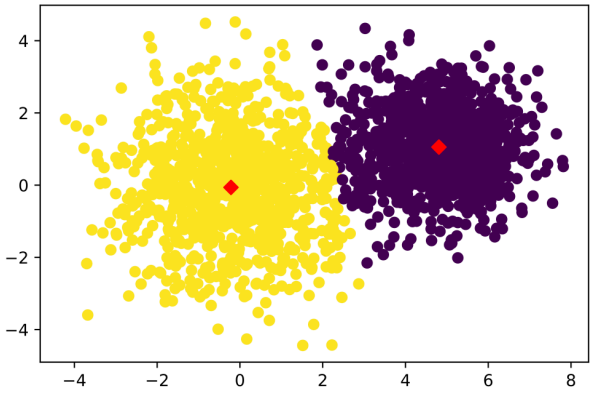
4. Evaluate the results for a range of K values (number of clusters) from K = 2, to K = 11 (but not all of these; try 3 different values, then try to pick 2-3 more that will help you find the best cluster outcome). When implementing the K-means algorithm (by itself and from your Spectral Clustering algorithm), you should try multiple initializations (say 5 to 10) and then select the 'best' result, for each K value. State how many different initializations you tried and why.

(1) The output should consist of a scatter plot of the data with the clusters differentiated by color, and with the centroids highlighted in a different color, size and shape than the corresponding cluster members.
(2) A plot of the distance metric, D, versus, K, for the best outcomes found for each of the K values you considered, for both K-means and Spectral approaches.
(3) Using the two best results from your K-means, that is, two different K-values that seem to give the best clustering result, use your GreedyKcenters algorithm to find an initialization and re-run K-means with this initialization. Describe how this does or does not improve your results.

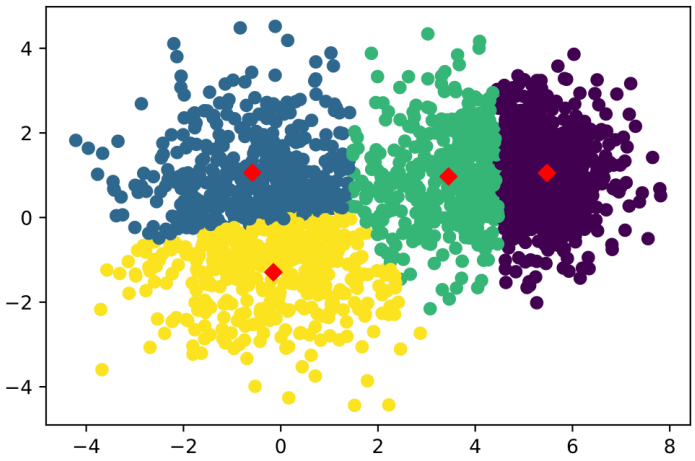
(1)
random initializations I implemented for KMeans are 10. Because the first series of centroids are randomly, arbitrarily selected. It can be anywhere even an outlier. So the more initializations, the more chance you get a correct clustering result.

KMeans only
Dataset 1

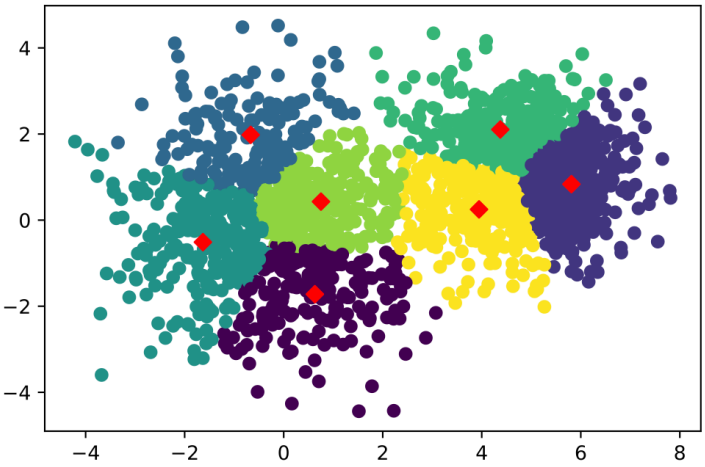
K = 2



K = 4



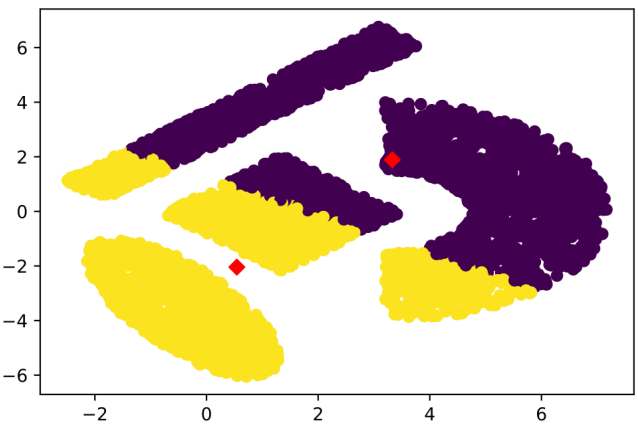
K = 7



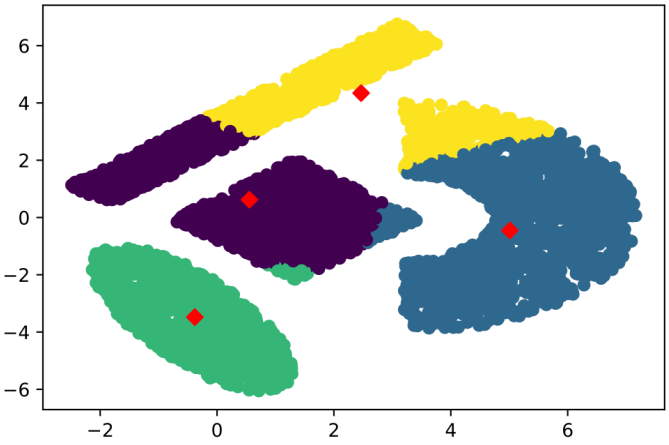
KMeans

Dataset 2

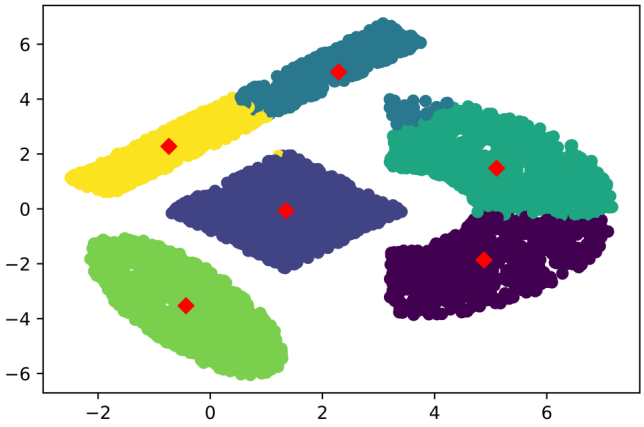
K = 2



K = 4

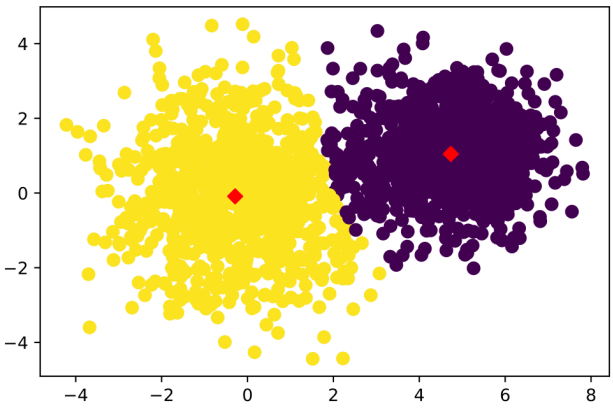


K = 6

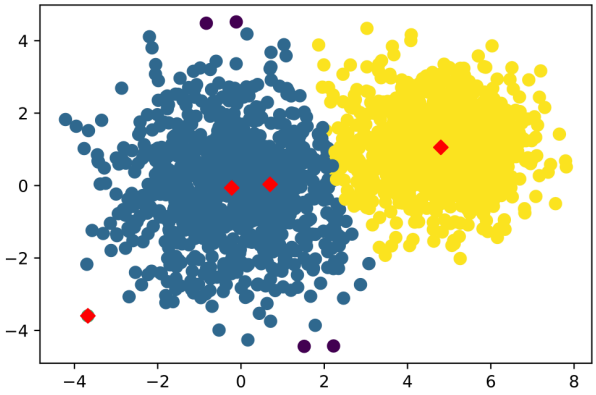


Spectral Clustering
Dataset 1

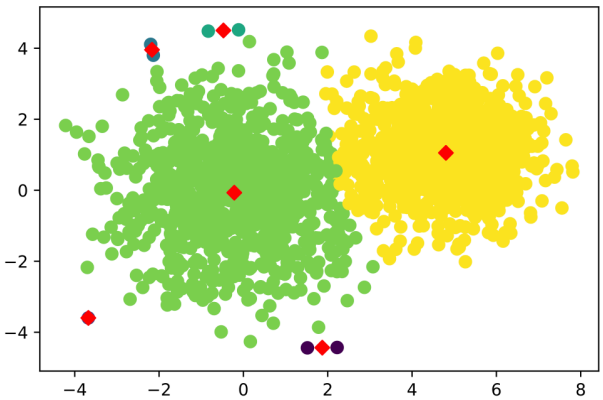
K = 2



K = 4



K = 6

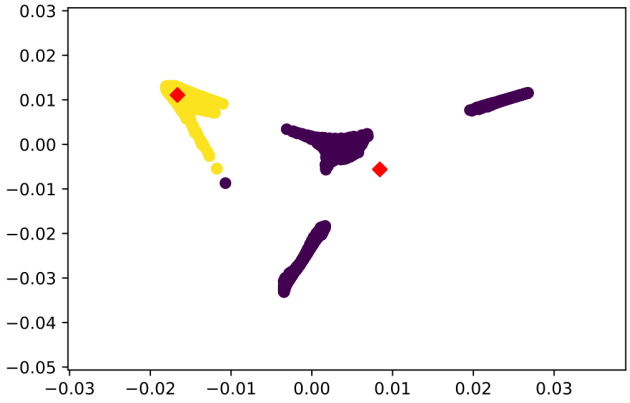


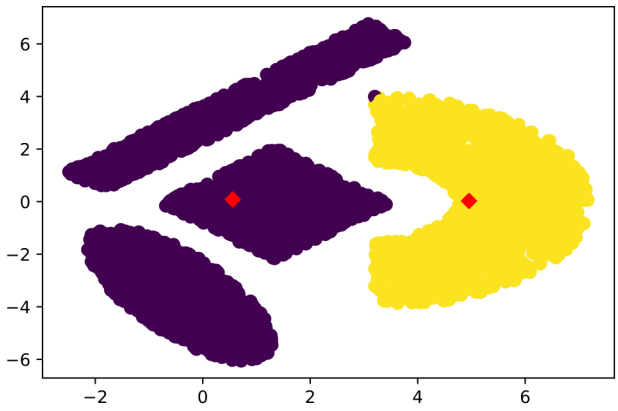
Spectral Clustering

Dataset 2

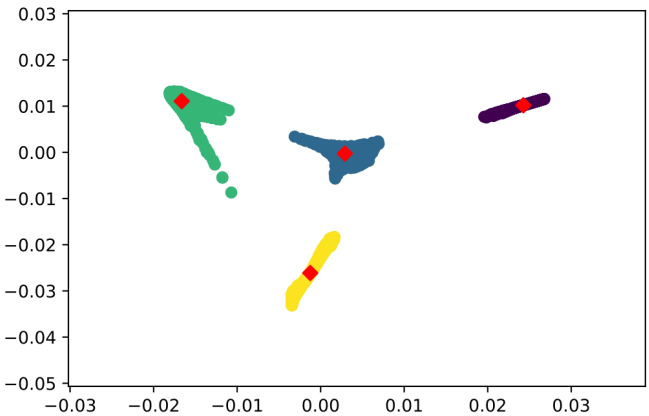
random initializations I implemented for KMeans with Spectral Clustering are 5. Because compute similarity matrix and perform decomposition is very time-consuming and have a high computational complexity.

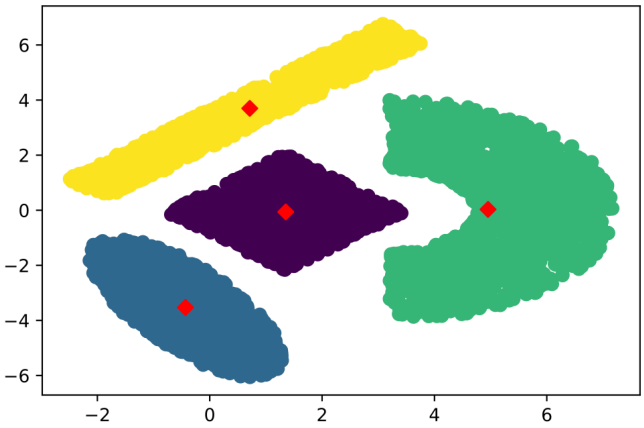
K = 2



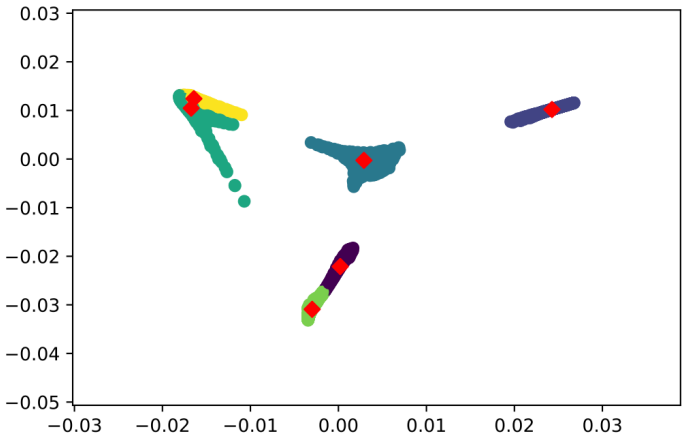


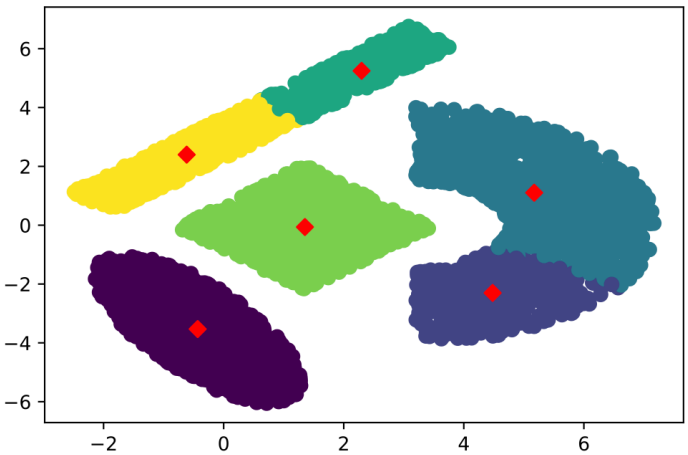
K = 4




6
4
2
0
−2
−4
−6
−2
0
2
4
6

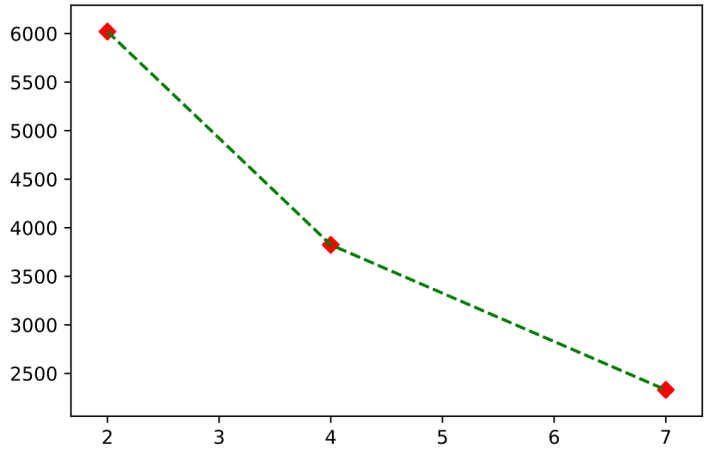
K = 6


0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
−0.05
−0.03
−0.02
−0.01
0.00
0.01
0.02
0.03

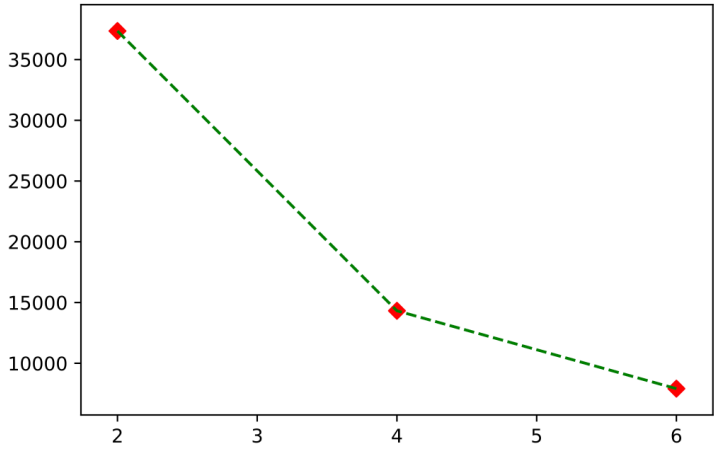

6
4
2
0
−2
−4
−6
−2
0
2
4
6

(2)

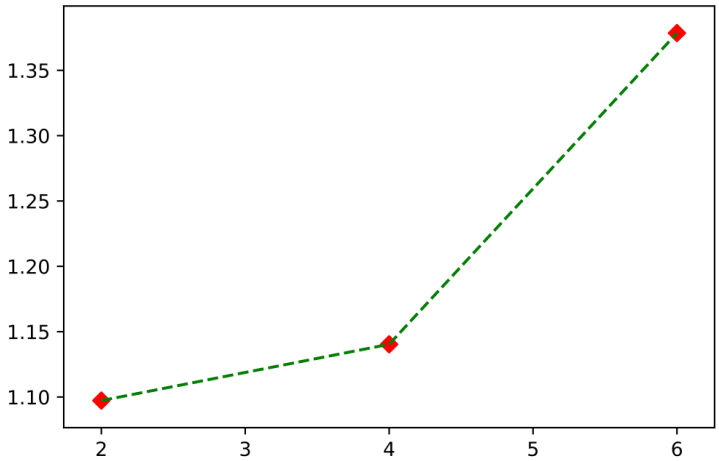
KMeans for Dataset 1



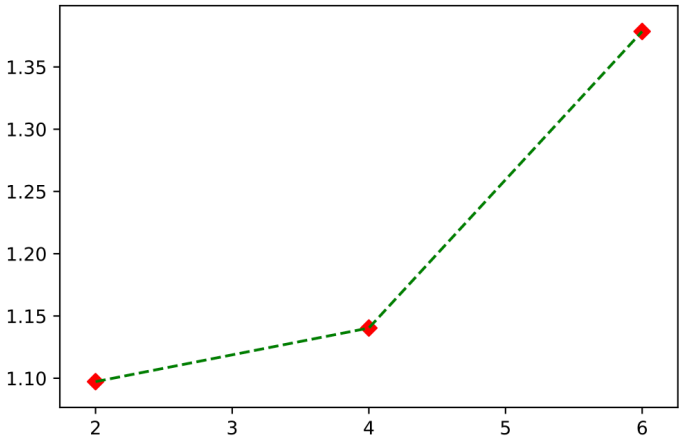
KMeans for Dataset 2



Spectral Clustering for Dataset 1

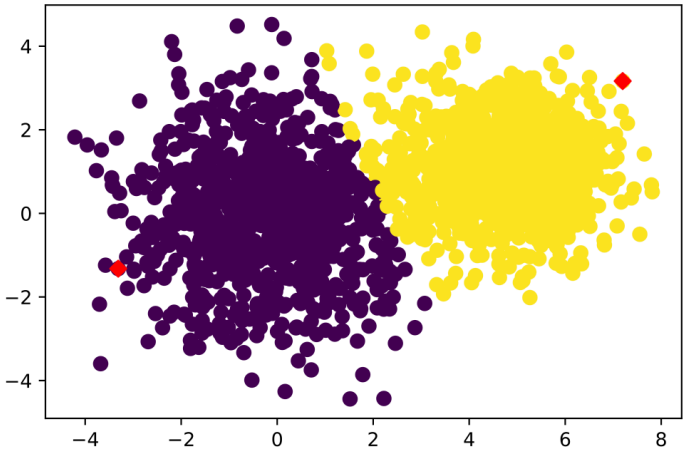


Spectral Clustering for Dataset 2

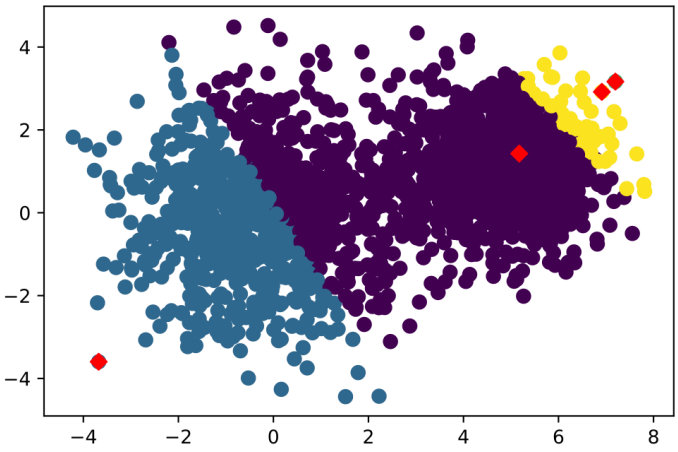


(3)

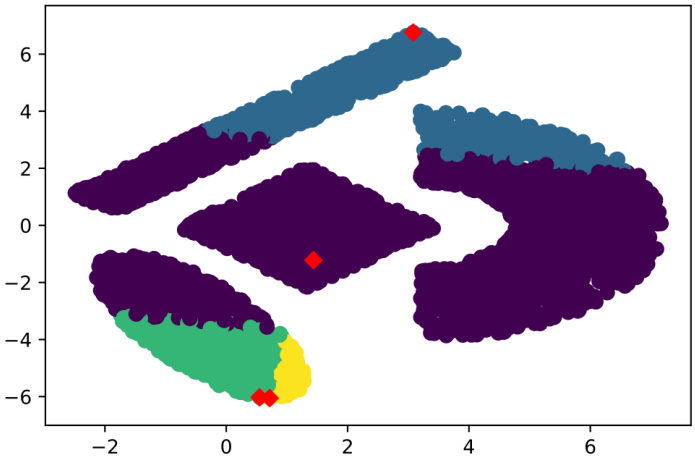
GreedyKCenters, K = 2, Dataset 1



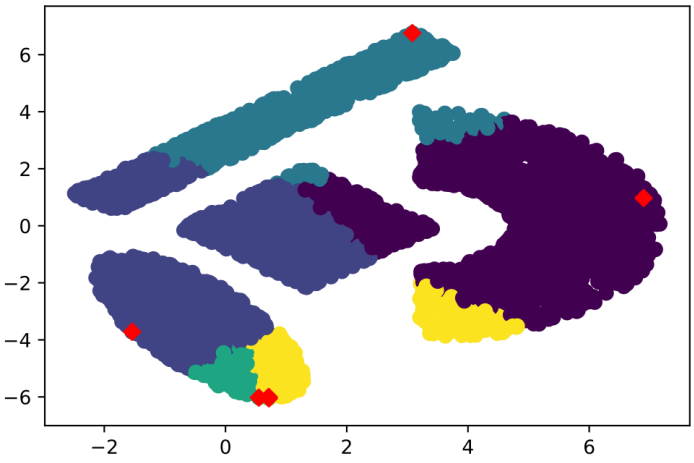
GreedyKCenters, K = 4, Dataset 1



GreedyKCenters, K = 4, Dataset 2



GreedyKCenters, K = 6, Dataset 2



Visually, the first centroid in on the edge of sample data. Those next centroids will be the farthest point from the first one (previous ones). So it does not improve the result with any of two datasets with either K = 2 or K = 4 or K = 6. Anyway, it does not help improve without swapping.

5. For the K-means algorithm, describe how you tested for convergence and why you choose this as a test.

The convergence condition I used is norm-based comparison of the distortion achieved $\| D_{p+1} - D_p \| < tol$. I chose *tol* to be 1×10^{-5} and 0.0001.

The distortion is defined by the the sum of the squared distances between each observation and its closest centroid. So if the distortion in *pth* iteration is very closed to the next iteration, like less than the threshold, we consider the specific number of centroids of this dataset has converged, to some extent. Considering the coordinates of all the data samples, the threshold is much less than the coordinates so if the difference of distortion is less than the threshold. We consider the iterations have converged.

6. Compare the results you get using your spectral clustering algorithm to the results obtained with your K-means algorithms for each of the data sets; you should compare the objective function values (D) and discuss qualitatively how well you think each clustering method worked (i.e., base this comparison on the results given in 4).

	Dataset 1		Dataset 2	
	KMeans	Spectral Clustering	KMeans	Spectral Clustering
K = 2	6020.257	1.112	37350.084	1.097
K = 4	3824.009	1.277	14318.542	1.140

For dataset 1, visually KMeans and Spectral Clustering both give awesome results. Also, we can see the objective function decrease from 6020 to 3824 when K increases from 2 to 4. KMeans works well on 'mixtures of Gaussians' as dataset 1.
For dataset 2, visually speaking, Spectral Clustering gives better result. Because there are four different shapes. KMeans just assigns the nearest centroids to each data point in dataset. Spectral Clustering compute the similarity matrix to evaluate how much similar within the data point.

7. Discuss how many "natural" clusters you think each data set has, and why.

Dataset 1: Two clusters
Dataset 2: Four clusters

Visually, sample data in first dataset is kind of mixed together. So intuitively there are two clusters in dataset 1. For dataset 2, those data points are shaped in specific geometric shapes. So there exist four clusters in dataset 2 obviously.

8. Provide a short analysis of the computational effort you found was required for each of the algorithms. State what measure you used to evaluate computational effort (e.g., running time or flops, or sth.), and how you observed this. Briefly

discuss how these computational efforts compare to your expectations.

Lloyd's Algorithm **O(nkd)** in each iteration: Computational complexity: for each centroid, you should compute the distance between sample data and centroids. The worst case will be $O(n^{(dk+1)}\log n)$.

Spectral Clustering you should compute the similarity matrix and eigenvalue decomposition $\mathbf{O(n^3)}$. When it comes to a large dataset. It is so expensive to compute the similarity matrix and eigenvalue decomposition, like dataset 2 in this assignment.

As the computational complexity of spectral clustering is O(n3). It is obvious when you run the code for spectral clustering, especially the two datasets contain so many observations. The process of computing similarity matrix and decomposition are both time-consuming.

9. Write a brief summary paragraph of your findings, that is, briefly summarize your KEY findings.

Spectral Clustering: Data points as nodes of a connected graph and clusters are found by partitioning this graph, based on its spectral decomposition, into sub-graphs.
K-means clustering: Divide the objects into k clusters such that some metric relative to the centroids of the clusters is minimized.

KMeans handle 'mixtures of Gaussian' perfectly with random initialization of centers. KMeans is ideal for discovering globular clusters like the ones shown below, where all members of each cluster are in close proximity to each other (in the Euclidean sense). Spectral clustering is more general (and powerful) because whenever K-means is appropriate for use then so too is spectral clustering (just use a simple Euclidean distance as the similarity measure). The converse is not true though. With Greedy-K-centers, sometimes it will have stuck at the 'outlier'.

Say if P data points each with N dimensions/features. Then using K-means you'll be dealing with an N by P matrix, while the input matrix to spectral clustering is of size P by P. You should now see the practical implications: spectral clustering is indifferent to the number of features you use (Gaussian kernel which can be thought of as an infinite-dimensional feature transformation is particularly popular when using spectral clustering). However, you will face difficulties applying spectral clustering (at least the vanilla version) to very large datasets (large P).